# Relational Mining
# for Compliance Risk

A Presentation for the Research Conference

Maury Harwood
Intelligent Business Solutions
NHQ Research
Maury.Harwood@irs.gov

David DeBarr
MITRE
debarr@mitre.org

# Suite of Tools

The effort will build a collection of analytical tools specifically designed to help the IRS explore relationships between taxpayers. Currently, Link Analysis is focused on flow-through relationships created by Partnerships, Trusts and Subchapter-S Corporations.

Tool development is being overseen by NHQ Research's Intelligent Business Solutions Group in collaboration with:

- Operating Divisions Research Staff
- MITRE Corporation
- NHQ Research Tax Return Database Group providing Compliance Data Warehouse Staff and Resources

# Objectives for Analysis of K-1 Tiered Transactions

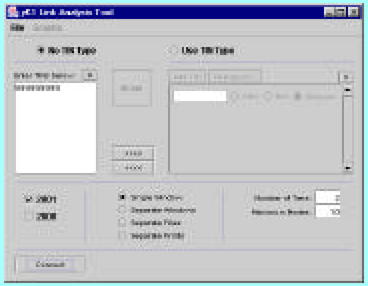
- Create flexible, easy to use tools that do not require users to be experts in data mining or specialized analytical techniques
- Identify business rules that select high interest networks
- Use automated techniques to distill high interest networks of K-1s down to manageable sets for analysis
- Achieve goals of IRS strategic plan
- Define and understand motivation for complex financial entities
- Begin to understand resource allocation needs

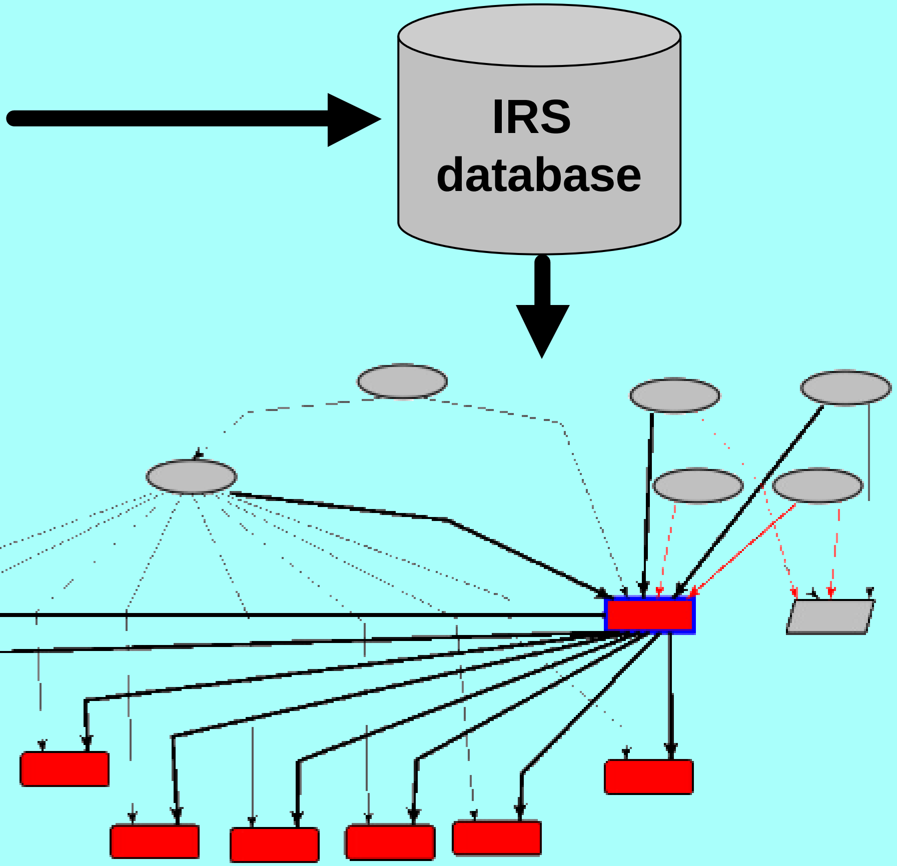
# Evolution of Effort

- IRS transcribes Schedule K-1 documents for the first time for tax year 2000
- Market review and technology assessment proposed using Link Analysis Technology (August 2002)
  - IRS Office of Research awarded a proof of concept contract to MITRE Corp, the IRS Federally Funded Research & Development Center
- Proof of concept for use of link analysis for flow-through entities demonstrated very quickly (May 2003)
- NHQ Research funds relational mining effort at MITRE through November 2004

# Pattern Visualization and Investigation, yK1

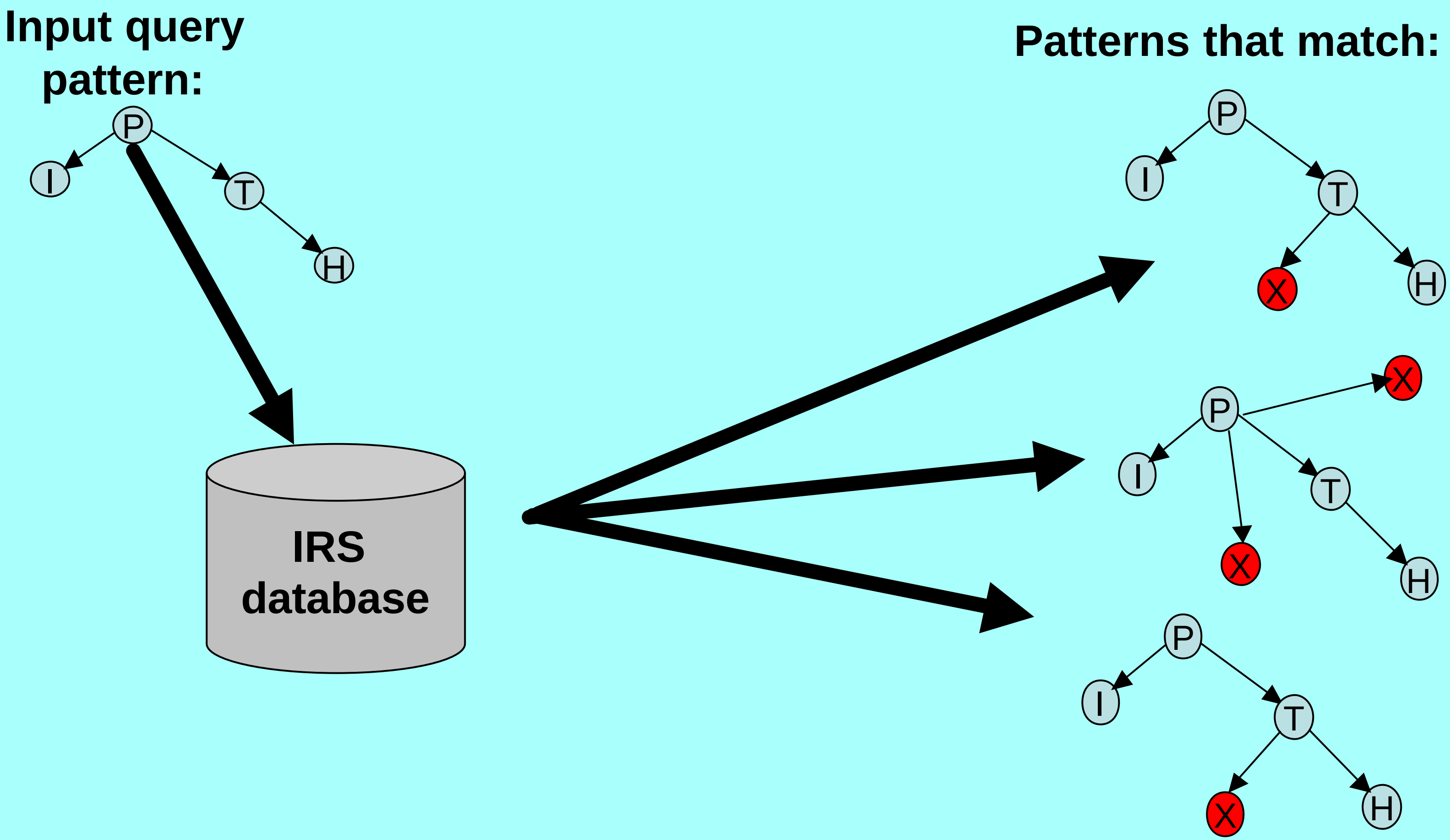
**Visualization Request for TIN:**





**Investment Structure Pattern of TIN and Related Entities:**

# Graph Query Pattern Matching (GQ)



# Status of Relational Mining

- Development of Specialized Data Structures and Algorithms
  - Requires iterative interviews with IRS Domain Experts
  - Customized model building and refinement
- Visualization tool (yK1) being actively tested by several users
- Graph Query Pattern Matching (GQ) Tool prototype is being evaluated
- Test cases delivered for classification/audit
- Several diverse domain areas and experts have been polled and recommendations for further development have been submitted by MITRE – Report to be completed
- Several research areas/algorithms to solve these problems are being actively pursued

# Current Efforts

- Economic Entity
- Temporal Differences
- Substructure Discovery
- Conceptual Clustering
- Learning – Active and SVM

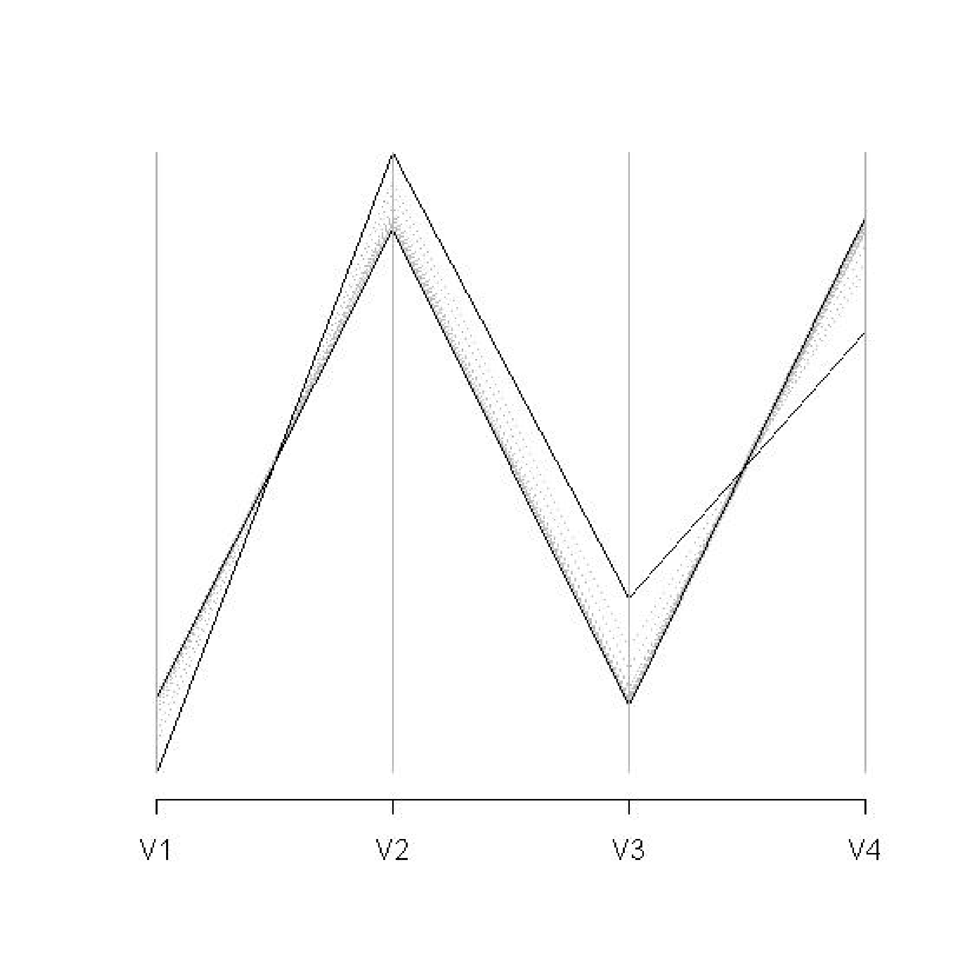
**All are being actively pursued**

# Detecting Abusive Transactions with Single-Class Support Vector Machines

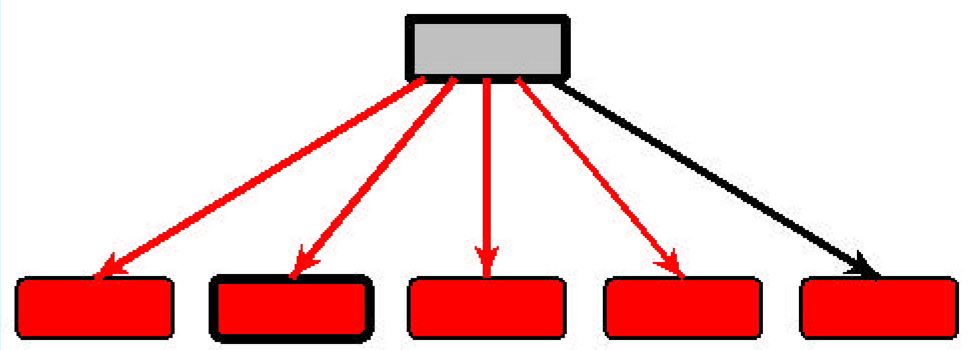
- Only requires abusive transactions to learn to recognize other abusive transactions
  - Known tax shelter examples can be provided as input
  - Remainder of the data can be analyzed to find similar behavior
  - Future returns can be analyzed to find similar behavior

- Employs quadratic programming to identify support vectors (positive class examples that define the optimal decision boundary)

# Parallel Coordinates Visualization of Data Describing 32 Abusive Transactions

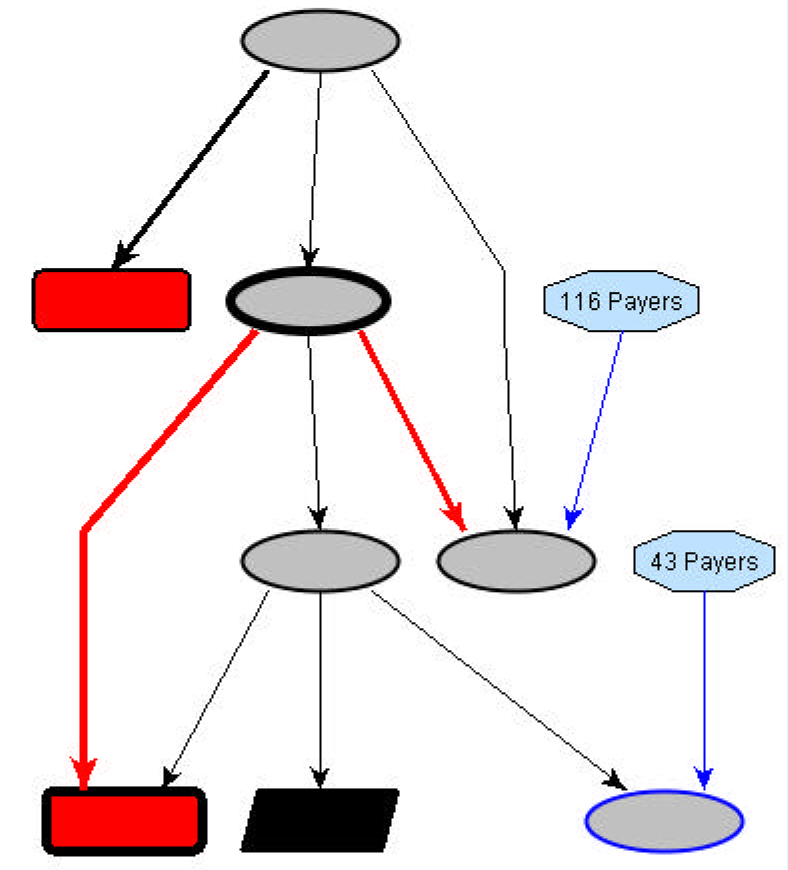


# Graph Illustrating a Support Vector (Support Vector = Prototypical Example)

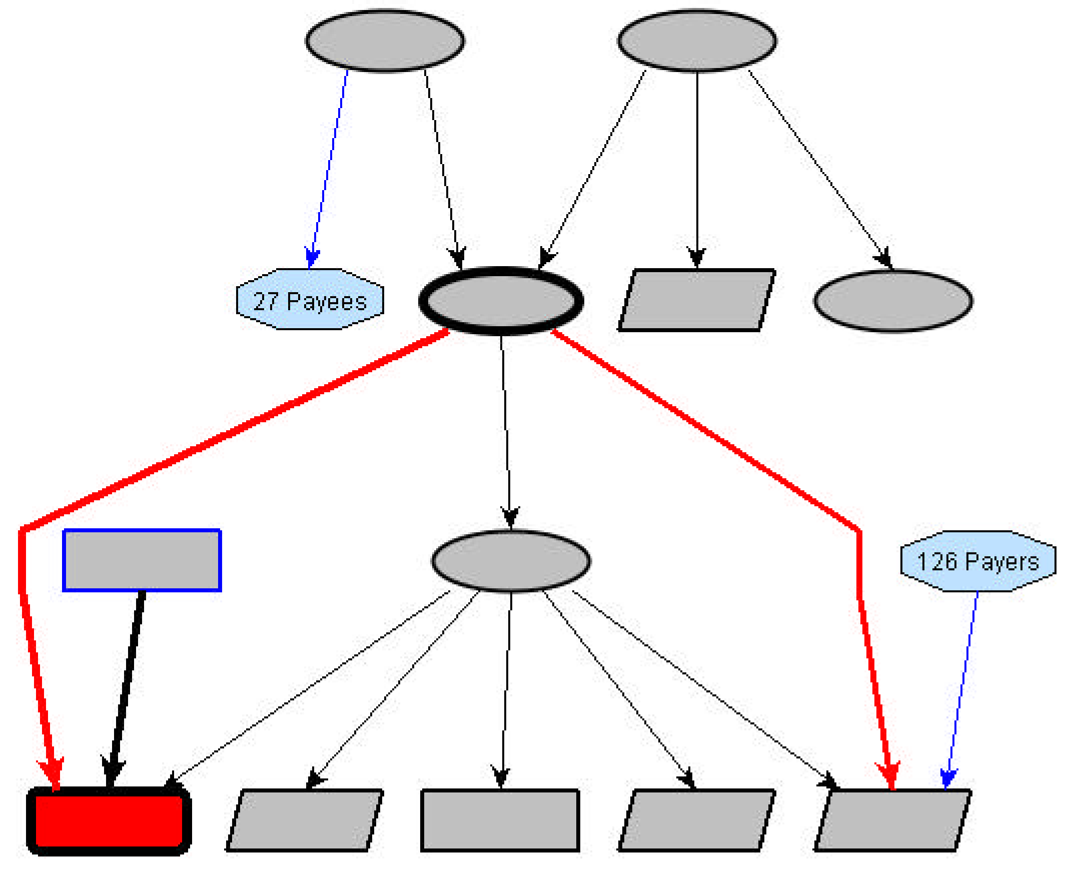


**Note: The black line is the result of omitting a minus sign during transcription. The black line was not characterized as part of the transaction.**

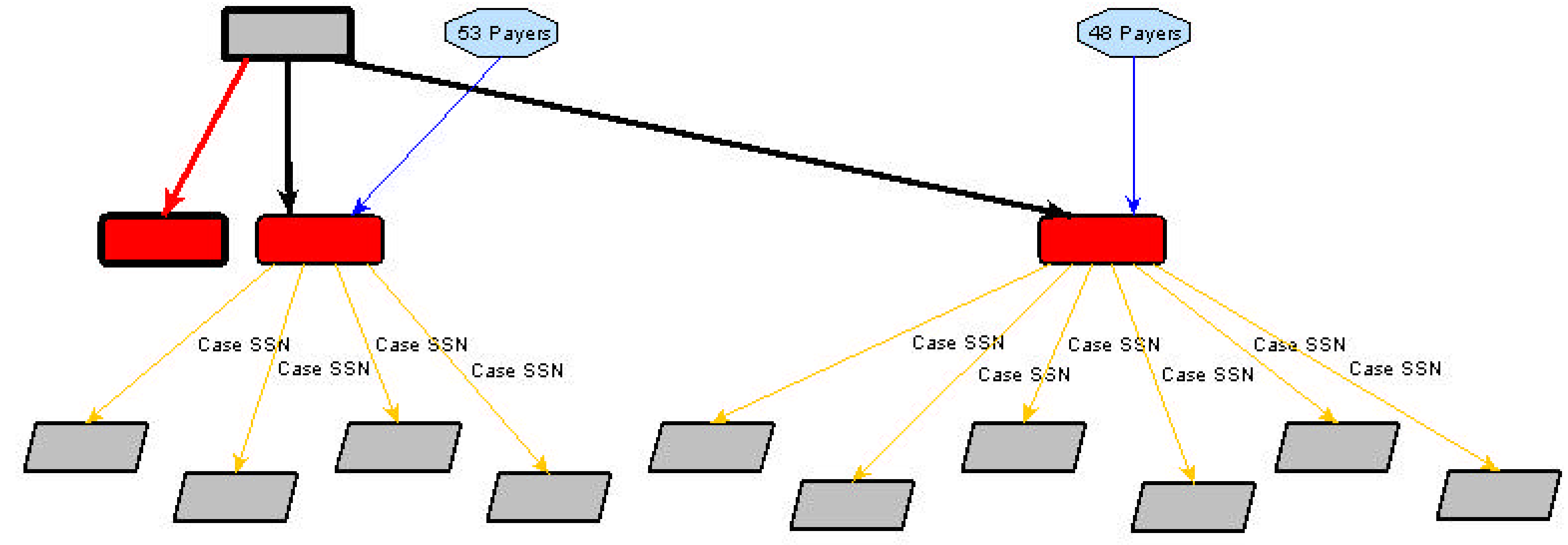
# Example of Abusive Transaction (found within same tax year)



# Example of Abusive Transaction (found within next tax year)



# Example of Different Type of Transaction (found within next tax year)



**Note: The black line is NOT the result of omitting a minus sign during transcription. This transaction was identified because of its similarity to the training data.**

# Active Learning

- Inactive learning
  - User provides large number of positive (abusive) and negative (non-abusive) transactions
  - Regression (or some other learning algorithm) is used to learn to distinguish positive examples from negative examples
  - Computer does not ask user to label any new examples
- Active learning
  - User provides small number of positive and negative transactions
  - Regression (or some other learning algorithm) is used to learn to distinguish positive examples from negative examples
  - Computer asks user to label a small number of new examples based on uncertainty
  - *This process allows the computer to* ***more quickly*** *develop a* ***better*** *decision boundary*

# Association Rules

- Schedule K-1 for Trusts, Partnerships, and S Corporations can be analyzed to identify unusual associations involving large dollar values

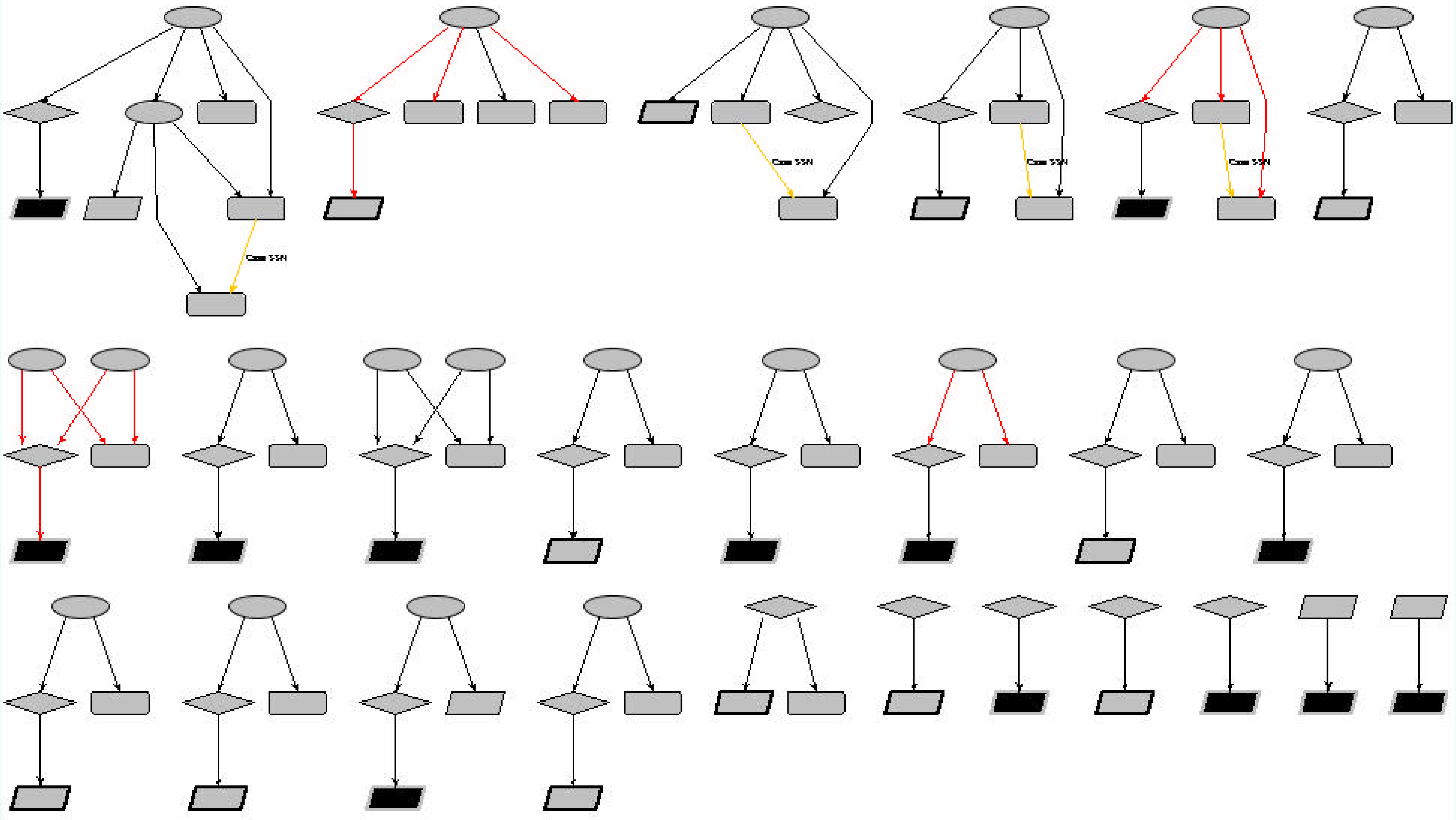
- Example

  IF Short Term Capital Gains = $100,000 AND Payee Country is XX (Specific Country)

  THEN NAICS Code is likely to be XXXXXX (Specific NAICS)

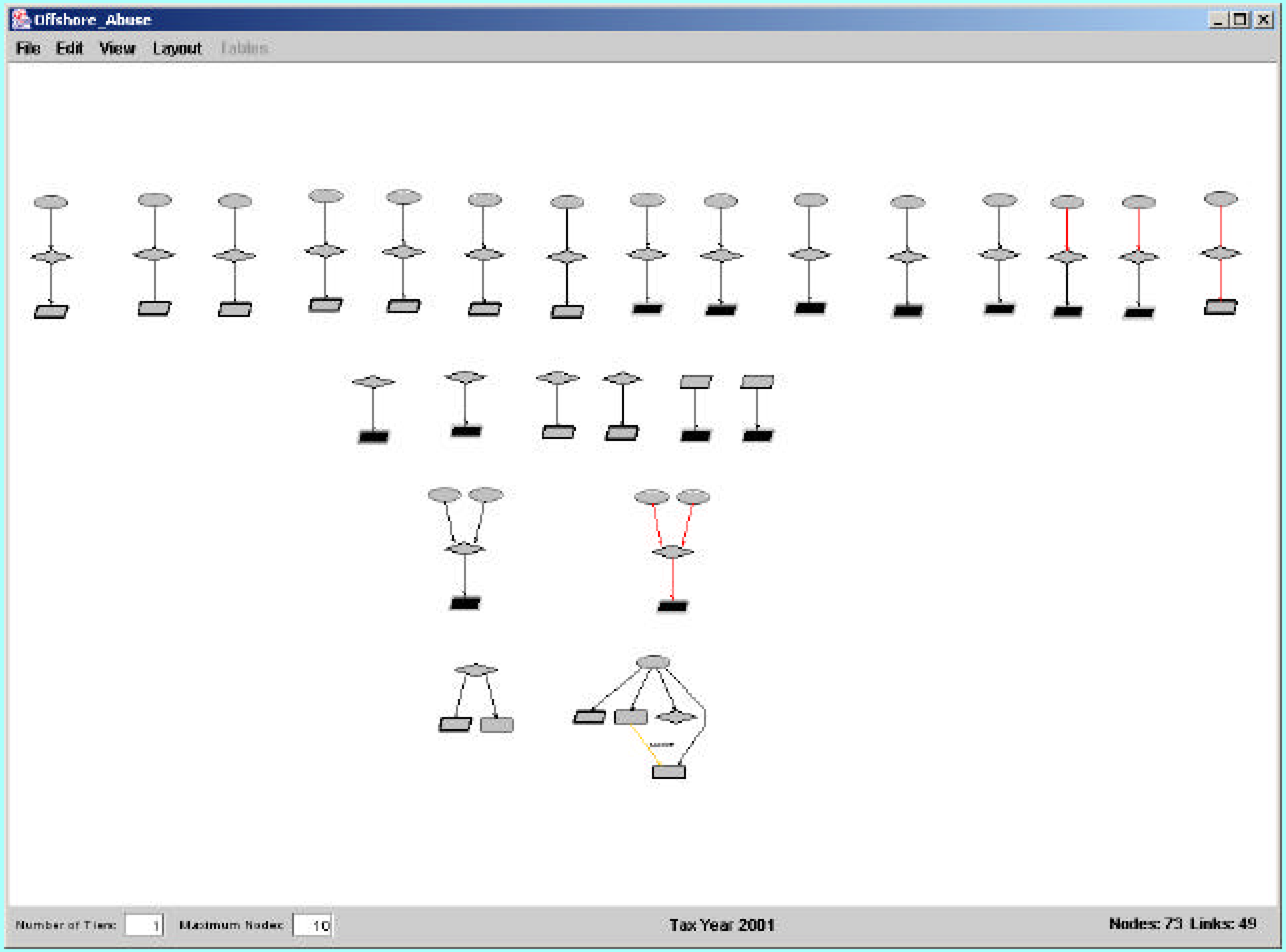
  - 24 K-1s indicated over $560,000,000 was allocated to a particular off-shore entity (address) during tax year 2001
  - No losses were allocated to the off-shore payees
  - There were over 250,000 K-1s with an off-shore payee for tax year 2001, yet these 24 transactions represented over 2.7% of all gains allocated to an off-shore payee

# Promoter Identified by Frequency Analysis
## (probability < 0.001 for shared values this frequent)



# Frequent Substructure Discovery and Conceptual Clustering

**These techniques can be used to summarize the data**



Number of Tiers: 1 Maximum Nodes: 10

Tax Year 2001

Nodes: 73 Links: 49